Beyond Information Retrieval—Medical Question Answering

Minsuk Lee[1], MS, James Cimino[2], MD, Hai Ran Zhu[3], MS, Carl Sable[4], PhD,
Vijay Shanker[5], PhD, John Ely[6], MD, Hong Yu[1], PhD
[1]Department of Health Sciences, University of Wisconsin-Milwaukee
[2]Department of Biomedical Informatics, Columbia University
[3]Department of Computer Science, Columbia University
[4]Department of Computer Science, Cooper Union
[5]Department of Computer Science, University of Delaware
[6]Department of Family Medicine, University of Iowa

Abstract *Physicians have many questions when caring for patients, and frequently need to seek answers for their questions. Information retrieval systems (e.g., PubMed) typically return a list of documents in response to a user's query. Frequently the number of returned documents is large and makes physicians' information seeking "practical only 'after hours' and not in the clinical settings". Question answering techniques are based on automatically analyzing thousands of electronic documents to generate short-text answers in response to clinical questions that are posed by physicians. The authors address physicians' information needs and described the design, implementation, and evaluation of the medical question answering system (MedQA). Although our long term goal is to enable MedQA to answer all types of medical questions, currently, we currently implement MedQA to integrate information retrieval, extraction, and summarization techniques to automatically generate paragraph-level text for definitional questions (i.e., "What is X?"). MedQA can be accessed at http://www.dbmi.columbia.edu/~yuh9001/research/MedQA.html.*

1. Introduction

The published medical literature and online medical resources are important sources to help physicians make decisions in patient treatment[1-3] and as a result, to enhance the quality of patient care[4, 5]. Although there are a number of annotated medical knowledge databases including UpToDate and Thomson Micromedex that are available for physicians to use, studies found that PubMed was still one of the resources most frequently used by physicians in large hospitals[3, 6]. Physicians often need to consult literature for the latest information in patient care[2, 7]. Information retrieval systems (e.g., PubMed) are frequently used by physicians. However, information retrieval systems frequently retrieve a vast amount of information in response to a specific user query. For example, querying PubMed about the drug *celecoxib* results in more than one thousand records. Physicians usually have limited time to browse the retrieved information. For example, studies found that physicians spend on average two minutes or less seeking an answer to a question, and that if a search takes longer, it is likely to be abandoned[1, 8-10]. Another evaluation study showed that it took an average of more than 30 minutes for a healthcare provider to search for answer from the PubMed, which means "information seeking is practical only 'after hours' and not in the clinical setting"[6].

Question answering is a rapid-developing technique that automatically analyzes thousands of articles to generate a short text, ideally, in less than a few seconds, to answer questions posed by physicians. Such a technique provides a practical alternative that allows physicians to efficiently seek information at point of patient care. This paper reports the research development, implementation, and a pilot evaluation of the medical question answering system (MedQA).

2. Background

The notion of communicating computer with humans in natural language started since the computer was invented. Earlier systems (e.g., ELIZA[11], PARRY[12], and SCHOLAR[13]) typically integrated simple rule-based approaches. Due to the complexity of human language, most of the early systems failed to be useful. Research advances in natural language processing have revived question answering developments, which have been largely driven by the Text REtrieval Conference (TREC) since 1999. The most recent TREC (2004) reported 77% accuracy for answering factoid questions (e.g., "How many calories are there in a Big Mac?") and 62.2% F-score for answer list questions (e.g., "List the names of chewing gums")[14]. Since 2003, TREC has provided evaluation to definitional questions (e.g., "What is a golden parachute?") that require long and complex answers. The Advanced Research and Development Activity (ARDA)'s Advanced Question & Answering for Intelligence (AQUAINT) program has since 2001 supported question answering techniques that generate long answers for scenario questions (e.g., opinion questions such as "What does X think about Y?")[15, 16]. Additionally, there are several question answering engines over the web (e.g., Brainboost[17]).

In contrast, fewer research groups are working on medical, domain-specific question answering. Zweigenbaum[18, 19] provided an analysis of the feasibility of question answering in the biomedical domain. Rinaldi and colleagues[20] adapted an open-domain question answering system to answer genomic questions (e.g., “where was spontaneous apoptosis observed?”) with the focus on identifying term relations based on a linguistic-rich full-parser. Niu and colleagues[21] and Delbecque and colleagues[22] also focused on term relation identification as a strategy for question answering. They combined shallow parsing with semantic information to assist term-relationship identification. Specifically, they mapped terms to the UMLS semantic classes (e.g., “Disease or Syndrome”) and then combined the semantic classes with surface cues or shallow parsing to capture term relations at the sentence level. For example, the word “plus” refers to the COMBINATION of two or more medications (e.g., “The combination of aspirin plus streptokinase significantly increased mortality at 3 months”). None of the systems[20-22], however, reported a fully-implemented question answering system that generates answers in response to users’ questions from a large text collection such as more than 15 million MEDLINE records.

In this paper, we report the first implemented medical question answering (MedQA) system that generates paragraph-level answers from both MEDLINE records and World-Wide-Web. We evaluated MedQA to answer definitional questions, i.e. questions with the form, “What is X?”.

Research in the context of open-domain definitional question answering have mainly focused on applying handcrafted lexico-syntactic patterns (e.g., “*<TERM>, ?(is|was)?Also?<RB>?called|named|known+as<NP>*”) to identify definitional sentences[23-25]. Similarly, Klavans and Muresan[26] extracted glossaries from medical text using a set of manually annotated surface cues (e.g., “also called”). In contrast to other systems[23-26], MedQA implements a set of lexico-syntactic patterns that are generated automatically. Additionally, MedQA is built upon four advanced techniques; namely, question analysis, information retrieval, answer extraction, and summarization techniques to generate a coherent answer to definitional questions. None of the previous systems[20, 21] reported the integration of all four techniques.

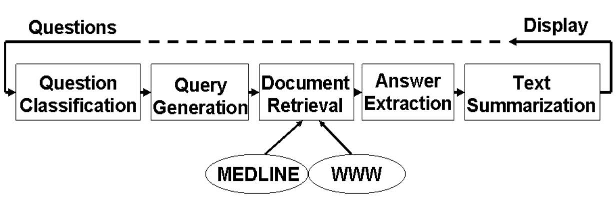

Figure 1 MedQA system components**.**

3. Methods

In this study, we develop natural language processing techniques, adapt existing natural language processing tools, implement, and evaluate MedQA. Figure 1 shows the MedQA overall system and the system components. *Question Classification* automatically classifies a question posed by a physician into a question type for which a specific answer strategy will be developed. *Query Generation* analyzes the question to extract noun phrases as the query terms. *Document Retrieval* applies the query terms to retrieve documents from either the Web documents or the locally-indexed MEDLINE collection. *Answer Extraction* automatically identifies the sentences that provide answers to questions. *Text Summarization* removes the redundant sentences and condenses the sentences into a coherent summary. The summary is then presented to the user who posed the question. In the following, we will first describe research development of each component, and then close with a pilot evaluation of the MedQA overall system.

Question Classification assigns a question posed by a physician to a specific category for which specific answer strategy is developed. Research has shown that medical questions can be classified by physicians into finite categories. For example, Ely and his colleagues created an "evidence taxonomy" to categorize medical questions. The “evidence taxonomy” incorporates five hierarchical categories [27]; namely, *Clinical* or *Non-clinical;* the *Clinical* questions are further divided into *General* versus *Specific*; *General* questions are divided into Evidence and *No-evidence*; and *Evidence* questions are divided into *Intervention* versus *No-intervention.* We explored supervised machine-learning approaches to automatically classify clinical questions into categories of the taxonomy created by Ely and his colleagues[28, 29]. Using a total of 200 annotated questions, our performance showed over 80% accuracy in 10-fold cross validation for classifying questions into the categories specified by the evidence taxonomy. In this study, we identify a total of 138 definitional questions[1] from the medical questions collected by Ely and colleagues[28, 29] and report the research development for answering definitional questions.

Query Term Generation and Document Retrieval applies LT CHUNK[30] to identify noun phrases from medical questions and then applies the noun phrases as the query terms to retrieve relevant documents. We apply the tool LUCENE[31] to index the MEDLINE collection, from which we retrieve relevant documents using the query terms. LUCENE takes Boolean and phrase queries and returns ranked documents based on the vector-space model[32], a TF*IDF based cosine similarity model that is used in information retrieval. To retrieve definitions that appear in the Web

[1] The 138 definitional questions can be accessed at http://www.dbmi.columbia.edu/~yuh9001/research/definitional_questions.htm.

documents, we use Google:Definition, the service that will be described in the next section.

Answer Extraction identifies from the retrieved documents relevant sentences that answer the questions. We automatically identified lexico-syntactic patterns, the patterns that incorporated both lexicon and syntax information, for identifying definitional sentences. Our strategy is to obtain an exhaustive list of lexico-syntactic patterns that has been generated from a set of definitional sentences. Specifically, we applied the Unified Medical Language System (UMLS 2005AA) terms (e.g., concepts and synonyms) as candidate definitional terms, and then identified their definitions with the *Google:Definition* service. *Google:Definition* provides definitions that seem to mostly come from web glossaries. A total of 36,535 UMLS terms (from the total of 1 million) had definitions specified by the *Google:Definition*; this corresponded to a total of 191,406 definitions; the average number of definitions for each definitional term is 5.2.

We then automatically identified lexico-syntactic patterns that comprise the definitional sentences. We applied a robust information extraction system Autoslog-TS[33] to generate automatically lexico-syntactic patterns. In the following, we will describe the AutoSlog-TS and how we applied it for lexico-syntactic pattern generation.

AutoSlog-TS is an information extraction system that automatically identifies extraction patterns for noun phrases by learning from two sets of unannotated texts. In our application, one collection of texts incorporates *relevant* or definitional sentences, and another collection of texts is *irrelevant* or *background* because it incorporates sentences that are randomly selected from the MEDLINE collection.

AutoSlog-TS first performs part-of-speech tagging and shallow parsing, and then generates every possible lexico-syntactic pattern within a clause to extract every noun phrase in both collections of *relevant* and *irrelevant* texts. It then computes statistics based on how often each pattern appears in the relevant texts versus the irrelevant texts and produces a ranked list of extraction patterns coupled with statistics indicating how strongly each pattern is associated with *relevant* and *irrelevant* texts. For example, one identified pattern is "*qureyterm, Formative Verb (e.g., "is" and "are"), Noun Phrase)*". We implemented the top 50 most frequent patterns captured by AutoSlog-TS into MedQA to capture definitional sentences.

Summarization techniques attempt to condense a stream of text into a shorter version while preserving its information content. MedQA builds on previous summarization and information retrieval techniques. It clusters sentences based on sentence similarity and selects the most representative sentence from each cluster. MedQA clusters sentences using the hierarchical clustering algorithms that have been evaluated in the biomedical domain[34]. To select the most representative sentence and then to generate a coherent summary, MedQA applies the centroid-based summarization technique (Radev et al., 2000). Specifically, MedQA first selects from each cluster one sentence that has the highest similarity to the rest of the sentences within the cluster. Then the selected sentences are ordered based on the similarity to the rest of selected sentences.

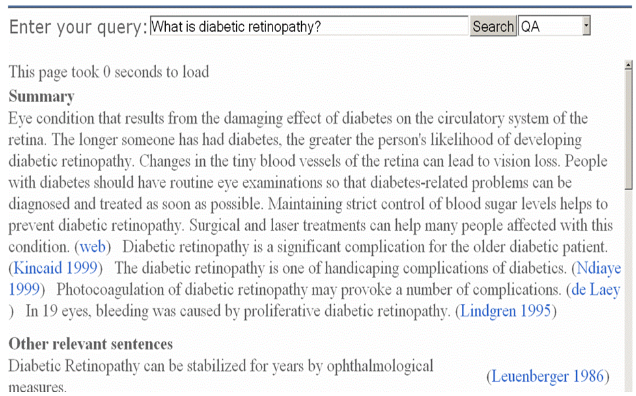

Figure 2: MedQA's output of the question "What is diabetic retinopathy?". A physician has given this answer the best score (i.e., "5") for *Answerability*. The parentheses provide links to the original documents from which the preceding sentences are extracted.

Web Definitions We have implemented *Google: Definition* to capture definitions from the Web documents. For each Web definition, we measure the similarity (i.e., TF*IDF) between the definition and the retrieved MEDLINE abstracts. We select a Web definition if it has the highest similarity score and yet the score is above an ad-hoc threshold we had defined.

Display The MedQA output a "summary" and "other relevant sentences". The summary incorporates a Web definition and followed by top five representative MEDLINE definitional sentences. "Other relevant sentences" displays the rest of MEDLINE definitional sentences. Figure 2 shows the MedQA's output in response to the question "What is diabetic retinopathy?". MedQA links each sentence to the original document resources, either on the Web, or the locally indexed MEDLINE collection.

4. Pilot Evaluation

We asked six physicians as evaluators to submit up to five self-generated definitional questions. Two physicians are the co-authors (JE and JC) of this paper. The four other physicians are on faculty in either hospitals or Universities and who are active in medical informatics research. The evaluators were asked to assign 1~5 scores that represent “very poor”, “poor”, “neutral”, “good”, and “excellent” to the following criteria:

Display: Which type of display you prefer? *Summary*, *Other Relevant Sentences*, or *Summary+Other Relevant Sentences*.

Answerability: Does the summary answer your question?

Sufficiency: Is the answer from the summary sufficient?

Coherency: Is the summary coherent?

Relevancy: Are sentences from the summary relevant?

Redundancy: Are sentences from the summary redundant?

Usefulness: Do the sentences from other relevant sentences incorporate useful content?

Coverage: Do the sentences from other relevant sentences incorporate a wide coverage of knowledge?

The six physicians posed a total of 21 questions. Our results show that no physicians prefer the display to be “other sentences” only. In 52.4% questions, physicians prefer “summary+other sentences”. Figure 3 shows the percentage for each possible response.

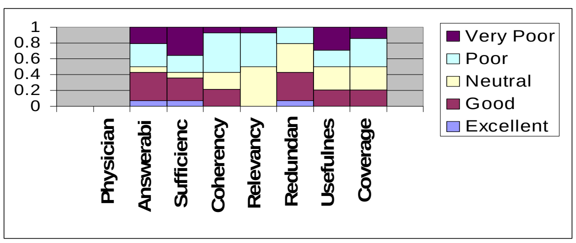

Figure 3 Evaluation Results

5. Software Environment

MedQA is implemented with Perl as the core platform and is running on a Macintosh PowerPC (dual 2 Ghz CPU, 2 GB of physical memory, Mac OS X server 10.4.2). The distributions of time spent among different components are 6 seconds for *Document Retrieval*, 6 seconds for *Answer Extraction*, and 6 seconds for *Text Summarization.*

6. Discussion and Future Work

Our evaluation results show that physicians scored MedQA high if the output incorporates a good answer, even if the good answer is mixed with some other inconsistent or irrelevant sentences. For example, A physician gave a best score (i.e., 5) for the answerability to the answer (as shown in Figure 2) of the question “What is diabetic retinopathy?”, even thought the output incorporates an irrelevant sentence (e.g., the sentence “In 19 eyes, bleeding was caused by proliferative diabetic retinopathy”).

Throughout the MedQA development, we identified a number of important research areas. Our current system implemented the shallow syntactic chunker LT CHUNK to capture noun phrases as query terms for answer extraction. However, we found LT CHUNK makes many mistakes. For example, LT CHUNK fails to identify “eating disorder” as the noun phrase in the question “What is eating disorder?”. The facts that LT CHUNK was trained on general English text, and not medical, domain-specific text, and that LT CHUNK was mostly trained on regular sentences, not questions have greatly undermined the capacity of LT CHUNK to efficiently capture noun phrases of medical questions. A comprehensive medical question answering system needs a robust and accurate parser that is specifically developed in the biomedical domain. Such a parser will also be useful for capturing lexico-syntactic patterns for answer extraction.

MedQA must be user-driven. For example, when a physician asks “What is the dawn phenomenon?” he wants to know not only the definition of this term, but also how to diagnose it and manage it. Essentially, a definition question (i.e., “What is X?”) requires answers beyond definitions (e.g., “what causes X?” and “How to treat X?”). One must work directly with physicians throughout the MedQA development.

Speed is an extremely important issue. Obviously, the higher is the speed, the greater is the user satisfaction. Question answering system incorporates many computationally intensive components (e.g., parsing and machine-learning), and further consume processing time because of the large number of documents that need to be processed; this makes it a challenge for a question answering system to deliver optimal response times with typical off-the-shelf hardware.

Semantic information plays an important role for both answer extraction and summarization and they are not captured in current MedQA implementation. Future work one shall capture the semantic information, identify statistical correlations, and combine the semantics with the lexico-syntactic patterns to efficiently identify sentences for answer extraction.

Although current MedQA’s capacity is limited: it only provides answers to definitional questions. It is our long-term goal to enable MedQA to answer other types of medical questions.

7. References

1. Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. Bmj. 1999 Aug 7;319(7206):358-61.
2. Straus S, Sackett D. Bringing evidence to the point of care. Journal of the American Medical Association. 1999;281:1171-2.

3. Cimino JJ, Li J, Graham M, Currie LM, Allen M, Bakken S, et al. Use of online resources while using a clinical information system. AMIA Annu Symp Proc. 2003:175-9.
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42.
5. Gosling AS, Westbrook JI. Allied health professionals' use of online evidence: a survey of 790 staff working in the Australian public hospital system. Int J Med Inform. 2004 May;73(4):391-401.
6. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P, et al. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. J Am Med Inform Assoc. 2002 May-Jun;9(3):283-93.
7. Sackett D, Straus S, Richardson W, Rosenberg W, Haynes R. Evidence-Based Medicine: How to practice and teach EBM. Edinburgh: Harcourt Publishers Limited; 2000.
8. Alper B, Stevermer J, White D, Ewigman B. Answering family physicians' clinical questions using electronic medical databases. . J Fam Pract 2001;50(11):960-5.
9. Jacquemart P, Zweigenbaum P. Towards a medical question-answering system: a feasibility study. Stud Health Technol Inform. 2003;95:463-8.
10. Takeshita H, Davis D, Straus S. Clinical evidence at the point of care in acute medicine: a handheld usability case study. Proceedings of the human factors and ergonomics society 46th annual meeting; 2002.
11. Weizenbaum J. ELIZA-A computer program for the study of natural language communication between man and machine. . Communications of the ACM. 1966(9):36-45.
12. Colby K, Weber S, Hilf F. Artificial paranoia Artificial Intelligence. 1971;2:1-25.
13. Collins A, Warnock E, Passafiume J. Analysis and synthesis of tutorial dialogues. The psychology of learning and motivation: Advances in research and theory; 1975.
14. Voorhees E. Overview of the TREC 2004 question answering track. NIST Special Publication.
15. Yu H, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP).
16. Bethard S, Yu H, Thornton A, Hatzivassiloglou V, Jurafsky D. Semantic analysis of propositional opinions. AAAI 2004 Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.
17. http://www.brainboost.com/ B.
18. Zweigenbaum P. Question answering in biomedicine. Workshop on Natural Language Processing for Question Answering, EACL 2003.
19. Zweigenbaum P. Question-answering for biomedicine: Methods and state of the art. MIE 2005.
20. Rinaldi F, Dowdall J, Schneider G, Persidis A. Answering questions in the genomics domain. ACL 2004 Workshop on Question Answering in Restricted Domain 2004.
21. Niu Y, Hirst G. Analysis of semantic classes in medical text for question answering. ACL 2004 Workshop on Question Answering in Restricted Domains.
22. Delbecque T, Jacquemart P, Zweigenbaum P. Indexing UMLS semantic types for medical question-answering. Connecting Medical Informatics and Bio-Informatics R Engelbrecht et al (Eds) ENMI 2005.
23. Liang L, Liu C, Xu Y-Q, Guo B, Shum H-Y. Real-time texture synthesis by patch-based sampling. ACM Trans Graph. 2001;20(3):127-50.
24. Blair-Goldensohn S, McKeown K, Schlaikjer A. Answering Definitional Questions: A Hybrid Approach. In: Maybury M, editor. New Directions In Question Answering: AAAI Press; 2004.
25. Cui H, Kan M, Cua T. Generic soft pattern models for definitional question answering. The 28th Annual International ACM SIGIR 2005.
26. Klavans J, Muresan S. Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. Proc AMIA Symp; 2001.
27. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. Bmj. 2000 Aug 12;321(7258):429-32.
28. Yu H, Sable C. Being Erlang Shen: Identifying answerable questions. Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasoning for Answering Questions 2005.
29. Yu H, Sable C, Zhu H. Classifying Medical Questions based on an Evidence Taxonomy. . AAAI 2005 Workshop on Qusetion Answering in Restricted Domains; 2005.
30. Mikheev A. Learning part-of-speech guessing rules from lexicon. Proceedings of COLING'96.
31. Goetz B. The Lucene search engine: Poweverful, flexible and free Available at http://luceneapacheorg/java/docs/. 2002.
32. Witten I, Moffat A, Bell T. Managing Gigabytes: Compressing and indexing documents and images.: Morgan Kaufmann Publishers. 1999.
33. Riloff E, Philips W. An introduction to the Sundance and AutoSlog Systems. : University of Utah School of Computing. ; 2004.
34. Lee M, Wang W, Yu H. Exploring supervised and unsupervised approaches to detect topics in biomedical text. . BMC Bioinformatics, Accepted. 2006.